

贝叶斯分类器笔记

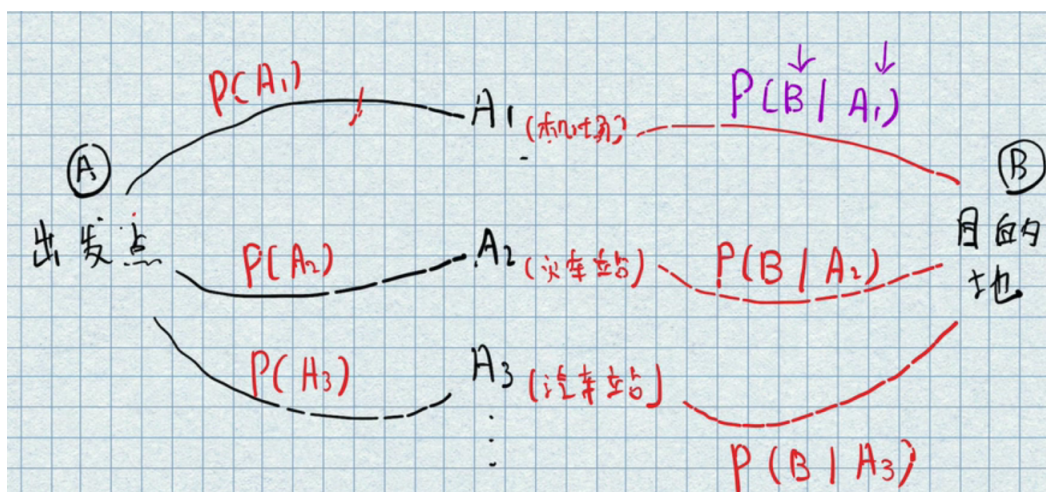
宋佳欢

2019 年 9 月 25 日

目录

1 全概率与贝叶斯公式	1
2 贝叶斯统计思想与贝叶斯估计	2
2.1 共轭先验分布	2
2.2 后验分布计算	2
3 朴素贝叶斯分类器	3
4 EM 算法	3
4.1 最大似然估计	3
4.2 高斯混合模型	3
4.2.1 EM 算法	4

1 全概率与贝叶斯公式



假设某人从 A 到 B 的路径有多条可以选择。则 A 到 B 的概率 P 为 (全概率公式):

$$p = p(A_1)p(B|A_1) + p(A_2)p(B|A_2) + p(A_3)p(B|A_3)$$

若已知到达了 B, 推断是从哪条路线过来的? 经过第一条路径到达的概率为 (贝叶斯公式):

$$p(A_1|B) = \frac{p(A_1)p(B|A_1)}{p}$$

【总结】

全概率公式：知因推果

贝叶斯公式：知果推因

2 贝叶斯统计思想与贝叶斯估计

两大学派：频率学派、贝叶斯学派。频率学派只利用利用样本信息（样本观测信息）和总体信息（总体分布或总体所属分布提供的信息）。

贝叶斯学派还利用了先验信息（在抽样之前有关统计问题的信息，主要时来源于经验、历史资料），对先验分布进行相应的加工，从而获得先验分布。根据总体分布、样本信息、先验分布，得到后验分布从而进行推断。

两个学派的本质区别：是否利用先验信息。

【注】总体依赖于参数 θ 的概率密度函数，在频率学派中记做 $P(x; \theta)$ ，而在贝叶斯学派中记为 $P(x|\theta)$ 。

2.1 共轭先验分布

定义：总体分布 $X \sim P(X|\theta)$ ，先验分布为 $\pi(\theta)$ ，假如抽样后算得的后验分布 $\pi(\theta|x)$ 与先验分布同属于一个分布族，则称先验分布 $\pi(\theta)$ 为总体分布 $X \sim P(X|\theta)$ 的一个共轭先验分布。

2.2 后验分布计算

条件概率：

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

贝叶斯公式：

$$P(x|\theta) = \frac{P(\theta|x)P(x)}{\sum P_i}$$

样本 X 与参数 θ 的联合概率分布：

$$h(X, \theta) = P(X|\theta) \cdot \pi(\theta)$$

边缘密度函数：

$$m(x) = \int_{\theta} h(X, \theta) d\theta = \int_{\theta} P(X|\theta) \cdot \pi(\theta) d\theta$$

后验分布：

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{P(x|\theta) \cdot \pi(\theta)}{\int_{\theta} P(X|\theta) \cdot \pi(\theta) d\theta}$$

贝叶斯估计：由后验分布来估计参数 θ

1. 使用后验分布密度函数做大作为点估计，即最大后验估计。

$$\max \pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{P(x|\theta) \cdot \pi(\theta)}{m(x)}, \quad \propto \max P(x|\theta) \cdot \pi(\theta)$$

$m(x)$ 为正则化因子

2. 后验分布中位数

3. 后验分布期望

3 朴素贝叶斯分类器

依据贝叶斯公式计算后验概率，将后验概率最大的作为该类别。

$$P(Y|X) = \frac{P(X|Y)\pi(Y)}{\sum P(X)}$$

1. 如果 $\pi(Y)$ 未知，则假设先验概率分布是等概率。所以有

$$\max P(Y|X) \rightarrow P(X|Y)$$

2. 如果 $\pi(Y)$ 已知: $\max P(X|Y)\pi(Y)$

求:

$$P(Y|X) \propto P(X|Y)\pi(Y)$$

根据条件概率公式:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

而联合概率分布未知，所以提出**类条件独立假设** (朴素之处): 即属性 x_i 之间不存在依关系。

$$P(X|Y) = \prod_{i=1}^n P(x_i|y_i)$$

4 EM 算法

4.1 最大似然估计

最大似然估计:

$$\operatorname{argmax} \left[\sum_{i=1}^N \log N(x_i | \mu, \varepsilon) \right]$$

分别对 μ 和 ε 求偏导，令其导数为 0，得到最优的参数。

4.2 高斯混合模型

高斯混合模型: 两个高斯模型混合的参数 $\bar{\theta} = \{\mu_1, \mu_2, \varepsilon_1, \varepsilon_2\}$ 。k 个高斯模型的表示

$$P(x|\bar{\theta}) = \sum_{l=1}^k \frac{1}{k} N(\mu_l, \varepsilon_l)$$

上式将每个高斯分布同等看待，即权重都是 $\frac{1}{k}$ ，使得组合的模型的概率分布的积分等于 1。但更常见的是多个模型不等权重组合的情况。得到高斯混合模型:

$$P(x|\bar{\theta}) = \sum_{l=1}^k \alpha_l N(\mu_l, \varepsilon_l) \quad (1)$$

$$\sum_{l=1}^k \alpha_l = 1 \quad (2)$$

$$\bar{\theta} = \{\mu_1, \mu_2, \dots, \mu_k, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_k, \alpha_1, \alpha_2, \dots, \alpha_k\} \quad (3)$$

在引入高斯混合模型之后，参数的最大似然估计变为：

$$\bar{\theta}_{MLE} = \operatorname{argmax}_{\theta} \left\{ \sum_{i=1}^N \log \left[\sum_{l=1}^k \alpha_l N(\mu_l, \varepsilon_l) \right] \right\}$$

如果再对每个参数进行求偏导并令为 0，这是很难做到的。因此需要使用迭代的算法。

定义下一轮迭代的参数和上一轮的参数之间的关系为：

$$\theta^{(g+1)} = \arg \max_{\theta} \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ$$

【注】 $P(X|\theta) = P_{\theta}(X)$ ，为两个写法。

其中 Z 为隐变量 (latent variable)， Z 的加入使得简化模型的解法。每个样本点 x_i 的隐变量 $z_i \in \{1, 2, \dots, k\}$ ，表示该样本点是属于哪个高斯分布的，因此将高斯混合模型简化为了单个的高斯模型。

但是 Z 的加入不能改变数据的边缘分布，即要使得：

$$P(x_i) = \int_{z_i} P_{\theta}(x_i|z_i) \cdot P_{\theta}(z_i) dz_i$$

那么 $P(z_i)$ 是根据高斯混合模型中的权重系数 α 来的，权重越大的高斯分布，其 $P(z_i)$ 也越大。

$$P(z_i) = \alpha_i$$

$$P_{\theta}(x_i|z_i) = N(\mu_i, \varepsilon_i)$$

因为 z_i 都是正整数，所以积分变为求和，仍等于高斯混合模型：

$$P(x_i) = \int_{z_i} P_{\theta}(x_i|z_i) \cdot P_{\theta}(z_i) dz_i = \sum_{i=1}^k \alpha_i N(\mu_i, \varepsilon_i)$$

4.2.1 EM 算法

EM 算法：

$$\theta^{(g+1)} = \arg \max_{\theta} \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ$$

怎么保证迭代算法的有效性，即每一次迭代之后，似然函数要越来越大，这样才是好的。

$$\log P(X|\theta^{(g+1)}) \geq \log(P(X|\theta^{(g)}))$$

所以要证明，加入了隐变量之后的迭代算法能够使得上式成立。根据条件概率公式：

$$\log P_{\theta}(X) = \log P_{\theta}(X, Z) - \log P_{\theta}(Z|X)$$

等式两边对分布 $P_{\theta^{(g)}}(Z|X)$ 取期望：

$$E[\log P_{\theta}(X)] = E[\log P_{\theta}(X, Z) - \log P_{\theta}(Z|X)]$$

$$\log P_{\theta}(X) = \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ - \int \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ$$

等式右边第一项就是迭代算法需要 argmax 的对象, 肯定是越来越大的, 因此只需证明等式右边第二项会逐渐递减, 则整个似然函数就会递增。令:

$$\int \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ = H(\theta, \theta^{(g)})$$

因此只需证明 $H(\theta^{(g+1)}, \theta^{(g)}) \leq H(\theta, \theta^{(g)})$, 就能证明算法的有效性。证:

$$H(\theta^{(g)}, \theta^{(g)}) - H(\theta, \theta^{(g)}) \geq 0$$

$$H(\theta^{(g)}, \theta^{(g)}) - H(\theta, \theta^{(g)}) = \int \log P(Z|X, \theta^{(g)}) \cdot P(Z|X, \theta^{(g)}) dZ - \int \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ$$